MEDIATOR BASED ARCHITECTURE TO ADDRESS DATA HETEROGENEITY

Rashed Mustafa[1], Hasan Hafizur Rahman[2]

[1]Department of Computer Science and Engineering University of Chittagong, Bangladesh
[2]Department of Computer Science and Engineering BGC Trust University Bangladesh

Email: rashed_mustafa78@yahoo.com, hasan_cse_cu@yahoo.com

Abstract: *Mediator Based Architecture is an emerging technology used to address data heterogeneity issues. As the number of data sources increases, the data integration process becomes an administrative and performance bottleneck because of data heterogeneity. Mediator-based approach allows the integration of data from heterogeneous data sources, which are usually not centralized. A mediator system is set between a number of data sources and applications. Now a day, the searching and combination of information from distributed, autonomous and heterogeneous software systems can be considered as a challenge for the computer users. The mainstream of this work is to address the issues of data heterogeneity where data are located in various distributed system in different format. Hence, multiple data accessing point is required to retrieve information from various distributed system. The mediator service allows single point of access, enabling the retrieving of appropriate information. To achieve this goal, the service dynamically integrates and customizes data from various data providers. This paper concentrates the investigation of data sources, interfaces and capabilities for mediators, and the design and implementation of a model- "**Integration Mediation Kit (IMK)".** The kit can think the entire source query in a scalable way.*

Keywords*: Data integration, mediators, queries processing, data heterogeneity, distributed database, Integration Mediation Kit.*

1. Introduction

The Mediator based architecture [1] works as a middleware, which acts as a link between different data sources. This work set-up a kit between user interface and the data source or other mediators for finding the required data set from differences in the various services and resources used by the applications. In this system, a person can quickly recover heterogeneous data from different sources. Mediator aims simplifying user interfaces by providing a uniform and transparent view of those services and resources that diverge, the reason for this is that, multi-vendors provide them, they have been developed in accordance with different protocols, or are used to support distinct application. Mediator is a generic term referring to a system layer of software that tries to overcome the heterogeneity problem. The main objective of this work is to find suitable information sources and enabling a remote system to process heterogenoeus data. The task supports an integrated, read-only, view of data that resides in multiple databases. A schema for the integrated view is available from the mediator, and queries can be made against that schema. The Integration Mediation Kit (IMK) effectively solves these problems by pointing suitable information sources and by accessing data from those information sources.

2. Related works

A number of technologies have been developed for integrating several data sources. It provides frameworks for composing, reconfiguring, and reusing systems. Mediator languages, module interactions and database interaction facilities have been developed. The idea of using mediator architecture for integrating heterogeneous data sources is well established in the present age of technology. The TSIMISS [3] and MIND [5] are also worked in the mediator field. TSIMISS integrates information in a common data model, Object Exchange model (OEM) [7], which can query by application. MIND is a CORBA [8] based multi-database system. The previous works mainly focus on the semantic heterogeneity for solving temporal semantic heterogeneity in evolving information systems, column heterogeneity as a measure of data

Date of submission : 02.06.2011 Date of acceptance : 07.05.2012

quality and peer-to-peer mediator solving technique.

3. The Concept of Mediator and Heterogeneity

More generally a mediator is a system that supports an integrated view [6] over multiple information sources. The goal of a mediation system is to give a user the ability to issue a single query that would access multiple sources to retrieve different pieces of the result and would assemble these pieces to provide a composite response to the query. A major problem sarises frequently when attempting to support information sharing among autonomous heterogeneous database systems is the occurance of representational differences that exist among related data objects in different component systems. The heterogeneity problems [4] at the data level appear when the requester and the provider of a service use different databases to conceptualize their domain. The mediator component achieves this by relying on a set of mappings between the source and target database .The task here is to support an integrated, read-only, view of data that resides in multiple databases. This work can solve the syntactic and structural heterogeneity by implementing the "Integrator Mediation Kit (IMK)". It is concerned about the removl of data heterogeneity from different data sources such as various vendor providing database systems, and general file systems.

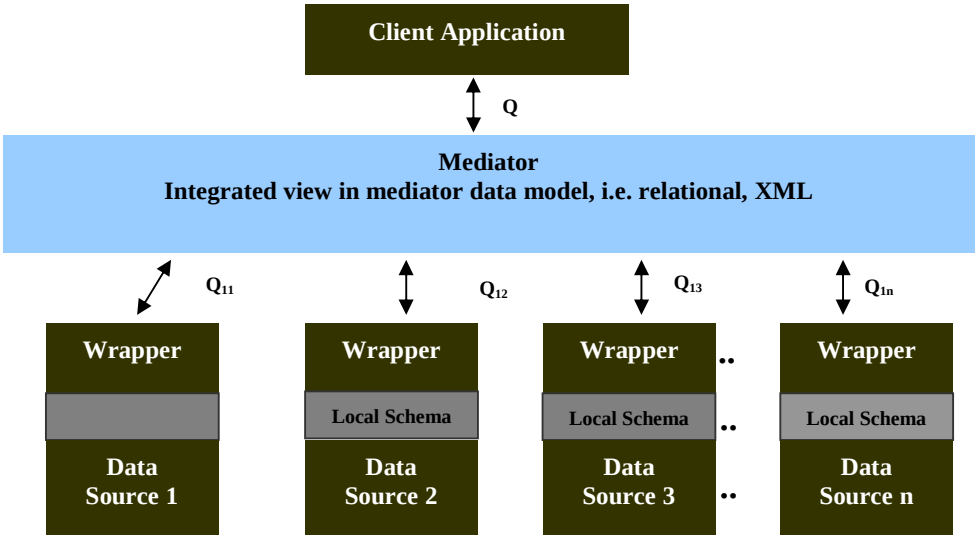

Figure 1: The Mediator / Wrapper Architecture

4. Methodology

The mediator based architecture is one of the most important that have been proposed to deal with the problem of integration of heterogeneous information. This architecture [2] consists of three parts, one is for client request, intermediate is for mediator layer and the other is for data sources. In this work the data sources are registered with the mediator. Registering these data sources can possible to retrieve the heterogeneous data from the multiple data sources. To solving the structural hetaerogeneity it is important to mappings, which are components that are used to perform structural conversion on data. This involves field names and data types. The filter part of this IMK can filter the requesting query into a specific and meaningful way as the user need. All data at the sources and access on per-need basis to retrieve and combine only the data that is relevant to a request. For that an intermediate software layer is introduced that presents to the users a logically integrated view of the data sources. This layer is a middle layer that separates the

functions related to data integration from the data management functions of the data sources and the presentation functions of the applications. The goal of this layer is to simplify, abstract, merge and explain data. It consists of mediator module defined as "Integration Mediation Kit".

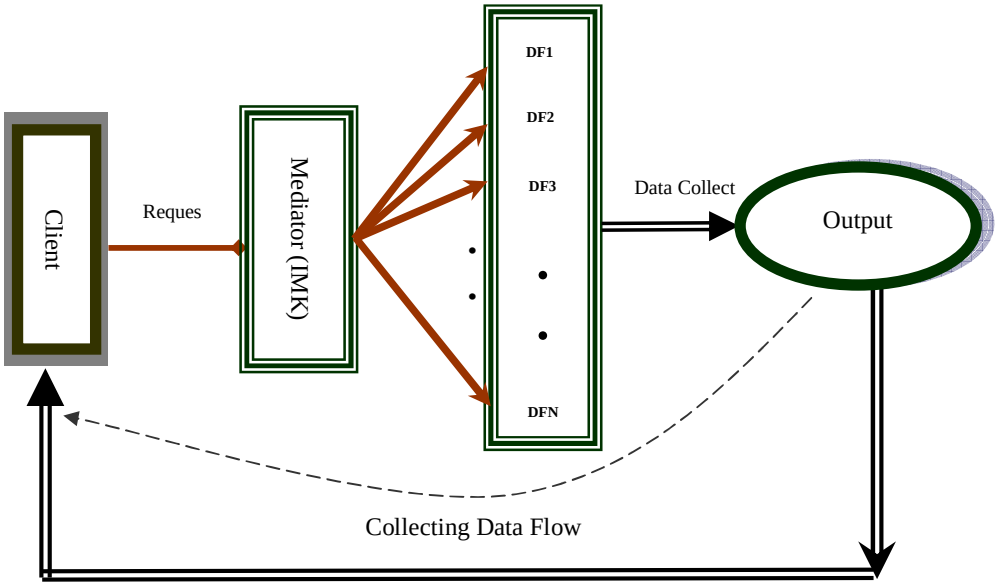

Figure 2: General View of Mediator system

The operation of the architecture is that the client request for a data (i.e. the authors publications against the authors name or authors id) then the requester sends this request to the mediator and the Integration Mediation Kit (IMK) is responsible for querying the data set, for this reason the IMK sends this query to other IMK in where the data resides. Finally all of these fragmented data are integrated into a common and responsive data set and sends as a return to the client. In the following general view of mediator system the first part is the client in where all of requesting is initialized, and then the requester sends this request to the mediator, and the mediator decides where the specific query result. Collecting the data from the data sources ensures the IMK and returns the data set to the client.

5. IMK Mediators

As described above, this mediator architecture encodes specific set of data and provides information for a higher level of applications. The information provided by the mediator is retrieved from underlying data sources. It is very easy situation that conflict arises so that it is solved by providing homogeneous interface to the clients. This architecture is an extension of the three-layer mediator architecture introduced in [8] four layer architecture is used for the IMK mediator run time environment. The four functional layers of the run time environment of a mediator describes in the following steps:

The application layer is the state where the client's data are stored; it is similar to all other application interface. The Data Integration layer is the collection of mediators, which are responsible for the integration of data those are provided by the local data server layer.

The Local Data Server responsible for storing the most commonly used data from the registering data sources and services provided by the local data source being integrated. The data source layer consists of the actual data sources such as vendor providing DBMS's raw files etc.

6. IIE and IMK

Interaction, Integration and Evolution (IIE) is the main strength of IMK. Interaction means how the entities are interrelated or making their relationship and hence becomes complimentary to each other (Integration). The complementary between the entities, make the system more dynamic that is evolutionary because interaction brings live to a system. For example, without the connection between input and output of an IMK (Integration Mediation Kit), will not enable it to perform its task. Hence the interaction is very important. Figure 2 demonstrates circular causation of IMK. Moreover, IMK has the characteristics that support IIE (Interaction Integration and Evolution) methodology.

7. Results

The IMK returns the information set from the heterogeneous data sources. For this architecture we choose the applicable domain in the multiple vendor providing data sources and the legacy file systems. Due to the different data sources involved, several data heterogeneity arises. Based on the specific field it is possible for finding the complete result by searching all of these registered data sources. It is also work well in the online bookshop where multiple data sources and legacy file systems exist. With existing multiple data sources, it is very common situation that heterogeneity also exists in those data sources. By finding the book name against the author's name, it can return the author's entire book set.

8. Conclusion

For other consideration as an example, If there are two database servers, one of them is in the one server another is in the other server. One database contains the author's id and author's name. In the same time other database stored additional information about author (author's publication date, author's publication name, author's journal name). The IMK helps one to get the full set of author's data by registering with the mediator system without establishing individual connection with the database server.

9. References

[1] *Gio Wiederhold.* "Mediators in the architecture of future information systems." IEEE Computer. Vol 25, No. 3. 1992.

[2] Bernadette Farias Lóscio, Ana Carolina Salgado, Vânia Maria Ponte Vidal: *Using Agents for Generation and Maintenance of Mediators*

[3] Garcia-Molina H. , Papakonstantinou Y. , Quass D. , Rajaraman A , Sagiv Y. , Ullman J. , Vassalos V., Widom J.. *"The TSIMMIS approach to mediation: Data models and Language,"* Journal of Intelligent Information Systems. 1997.

[4] *Integrating Heterogeneous Data Sources using the COIL Mediator Definition Language* Christian Och, Roger King, Richard Osborne.

[5] Dogac A, Dengi C, and Öszu. M.T "*Distributed Object Computing Platforms*." Communications of the ACM. September 1998, Vol. 41. No. 9

[6] Mariella Di Giacomo, Mark Martinez, and Jeff Scott: *A Large-Scale Digital Library System to IntegrateHeterogeneous Data of Distributed Databases, 1997.*

[7] Roy Goldman, Sudarshan Chawathe, Arturo Crespo, Jason McHugh. "*A Standard Textual Interface Format for the Object Exchange Model (OEM).*" Stanford Technical Report. 1996.

[8] The Object Management Group: *The Common Objec Request Broker: Architecture and Specification*, Revision 2.2, http://www.omg.org, February 1998

Biography

Rashed Mustafa is serving as an Assistant Professor (study leave) in the department of Computer Science and Engineering, University of Chittagong, Bangladesh and also he is pursuing PhD in Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Graduate University of Chinese Academy of Sciences Beijing, China.

Hasan Hafizur Rahman is serving as a lecturer in the department of Computer Science and Engineering, BGC Trust University Bangladesh and also he is pursuing MS in the department of Computer Science and Engineering, University of Chittagong, Bangladesh.